nature communications

Article

https://doi.org/10.1038/s41467-023-38795-w

A computational method for cell type-specific expression quantitative trait loci mapping using bulk RNA-seq data

Received: 21 April 2022

Accepted: 16 May 2023

Published online: 25 May 2023

Check for updates

Paul Little[1] ✉, Si Liu[1], Vasyl Zhabotynsky[2,3], Yun Li[2,4], Dan-Yu Lin[2,3] & Wei Sun[1,2,5] ✉

Mapping cell type-specific gene expression quantitative trait loci (ct-eQTLs) is a powerful way to investigate the genetic basis of complex traits. A popular method for ct-eQTL mapping is to assess the interaction between the genotype of a genetic locus and the abundance of a specific cell type using a linear model. However, this approach requires transforming RNA-seq count data, which distorts the relation between gene expression and cell type proportions and results in reduced power and/or inflated type I error. To address this issue, we have developed a statistical method called CSeQTL that allows for ct-eQTL mapping using bulk RNA-seq count data while taking advantage of allele-specific expression. We validated the results of CSeQTL through simulations and real data analysis, comparing CSeQTL results to those obtained from purified bulk RNA-seq data or single cell RNA-seq data. Using our ct-eQTL findings, we were able to identify cell types relevant to 21 categories of human traits.

Studying the variation of gene expression is essential for understanding cellular and molecular biology. Gene expression can vary significantly across different cell types, and that the composition of cell types can vary across tissue samples[1]. As a result, variation in gene expression observed in bulk tissue samples can be due to both cell type-specific expression and variations in cell type compositions[2]. Investigating gene expression quantitative trait loci (eQTLs), or genetic variants associated with gene expression, is a powerful approach for studying the genetic basis of complex traits[3,4]. Several recent studies found that many genetic loci implicated in human diseases are associated with certain cell types[5–8]. By studying cell type-specific eQTLs (ct-eQTLs), we can gain further insights into the genetic basis of complex traits[9–11].

A popular method to study ct-eQTLs using bulk tissue gene expression data is to include an interaction between the genotype at a genetic locus and the abundance of a cell type in a linear model[3,4,11–17]. However, linear models require the residual variation in gene expression to be constant across samples. Therefore, it is often necessary to use a log transformation or normal quantile transformation of RNA-seq count data to stabilize variance. These transformations can result in nonlinear relationships between transformed gene expression and cell type proportions, leading to a mis-specified linear model. An alternative and more appropriate modeling approach is to use negative binomial regression to directly model RNA-seq count data. In addition to total read count (TReC), RNA-seq data can also provide information about allele-specific expression (ASE). By incorporating both TReC and ASE, it is possible to increase the power of eQTL mapping by taking advantage of the allelic imbalance of gene expression caused by *cis*-acting eQTLs. We have developed a method called TReCASE that uses this approach[18,19]. Most local eQTLs are *cis*-eQTLs, and the terms are often used synonymously. In this paper, we use the term "*cis*-eQTL" to refer specifically to *cis*-acting eQTLs that lead to allelic imbalance.

We have previously developed a method called pTReCASE for eQTL mapping using RNA-seq data from tumor samples, where we treated tumor and non-tumor cells as two distinct cell types with known composition[20]. However, this approach is limited to situations

[1]Biostatistics Program, Public Health Science Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [4]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Department of Biostatistics, University of Washington, Seattle, WA, USA. ✉e-mail: plittle@fredhutch.org; wsun@fredhutch.org

where there are only two cell types and where cell type proportions vary significantly across samples. It treats bulk TReC as the sum of TReC from the two cell types. In more general situations with an arbitrary number of cell types, there are several challenges to ct-eQTL mapping. For example, a cell type may have nearly constant proportions across samples, which can make it difficult to accurately estimate the ct-eQTL effect for that cell type. Additionally, a gene's expression may be zero or very low in some cell types, making it difficult or impossible to estimate eQTL effects in those cell types. In this paper, we have designed a flexible and robust computational framework from scratch to handle these challenges.

Although single cell RNA-seq (scRNA-seq) data have become more widely available and can be used to study ct-eQTLs, there are still some limitations. First, scRNA-seq is expensive for studies with large sample sizes, and it also requires high quality samples. Additionally, scRNA-seq may not provide a representative sampling of all cell types in a tissue sample, and the inherent sparsity of scRNA-seq data can make it difficult to accurately assign cell types to individual cells. Our method allows for a new study design: collecting scRNA-seq data from a subset of samples, along with bulk RNA-seq data from all samples. The scRNA-seq data can be used to create a cell type-specific gene expression reference, and then the bulk RNA-seq data can be used for ct-eQTL mapping after estimating cell type proportions using the reference.

Results

A brief introduction of CSeQTL and OLS method

Our method is called cell type-specific eQTL or CSeQTL for short. CSeQTL jointly models total read count (TReC) and allele-specific read count (ASReC) as a function of covariates, cell type composition, and the genotype at a single nucleotide polymorphism (SNP). More specifically, TReC and ASReC are modeled by a negative binomial and a beta-binomial distribution, respectively, with shared parameters for genetic effects[18]. Unlike the TReCASE and pTReCASE methods, CSeQTL is designed to handle challenging situations where cell type-specific gene expression may be zero or very low, or the proportion of one or more cell types may be close to zero or lack variation. These challenges can make it difficult or impossible to accurately estimate eQTL effects. We address these issues by using several computational solutions, including trimming outliers of TReC to increase the robustness of our estimates and iteratively detecting and removing non-expressed cell types.

We compare CSeQTL to a linear model approach that we refer to as the ordinary least squares (OLS) method. To implement the OLS method, we first apply an inverse normal quantile transformation to read-depth normalized TReC for each gene. Next, we define a reference cell type (usually the one with the highest average abundance) and fit a linear model with transformed gene expression as the dependent variable and the following covariates as independent variables: the proportions of all cell types except the reference cell type, the genotype at a SNP, and the interactions between the SNP genotype and the proportion of each non-reference cell type. Other covariates such as age, sex, and batch can also be included. With this model setup, the ct-eQTL effect for the reference cell type is the main effect of the SNP genotype, and the ct-eQTL effect for a non-reference cell type is the sum of the genotype's main effect and the effect size of the corresponding interaction term. This model is the same as the one used by Aguirre-Gamboa et al.[16].

CSeQTL controls type I error and has much higher power than OLS

We conducted simulations to evaluate type I error and power of CSeQTL in a variety of settings. First, we varied the baseline expression (i.e., gene expression of the reference allele) across cell types. Second, we considered three scenarios of cell type composition variation for three cell types, referred to as CT1, CT2, and CT3 (Fig. 1a). In scenario 1, cell type proportions were generated independently and identically distributed, and then normalized to sum to one. This scenario represents an ideal, but unrealistic, situation. In scenario 2, we created more realistic cell type proportions by setting the average abundance of CT3 to be lower than CT1 and CT2, and by reducing the variance in the proportion of CT3. This scenario represents a more difficult situation for ct-eQTL mapping of CT3. In scenario 3, we added outlier proportions to the simulated proportions of scenario 2 to mimic observations in real data. We also conducted a secondary set of simulations to explore the performance of CSeQTL given noisy estimates of cell type proportions (Supplementary Fig. 1).

We set the mean expression of the reference allele in CT1 be 500. For example, if the reference/alternative allele is A/T, then the mean expression in an individual with genotype AA is 1000. We set the ASReC to be 5% of the TReC. We also set the fold change for the reference allele gene expression of CT2 or CT3 vs. CT1 to be 0.1, 1.0, or 10. Following the TReCASE model, TReC and ASReC were simulated conditional on phased SNP genotypes, cell type proportions, expected expressions per allele and cell type, and other covariates. The sample size was 300. All the eQTLs were set to be *cis*-eQTLs that influenced both TReC and ASReC and we specified eQTL effect size by fold change of alternative allele B vs. reference allele A. In a global null situation, all ct-eQTL effects were set to be 1.0 (Fig. 1b). In another mixed null/alternative situation, we allowed CT1's eQTL effect to vary from $\exp(-1)$ to $\exp(1)$, set the eQTL effect for CT2 to be 1 (i.e., no eQTL effect), and set the eQTL effect for CT3 to be 1.5. This design allowed us to assess power in CT1 and CT3 and type I error in CT2 simultaneously (Fig. 1c).

Under both global null and mixed null/alternative situations, CSeQTL controls type I error but OLS has apparent type I error inflation in several configurations (Fig. 1b, c). Focusing on the mixed null/alternative situation, we found that under scenario 1, when the three cell types have the same distribution of proportions, CSeQTL generally has higher power than OLS. When the baseline expression of CT3 is low (0.1 fold of CT1), OLS's power in CT3 is positively correlated with CT1's eQTL effect size even though CT3 has a constant effect size throughout. This "leaking" of eQTL effect from CT1 to CT3 is likely due to the transformation of gene expression. OLS also suffers from inflated type I error (i.e., eQTL findings from CT2) in cases where CT2 has lower baseline expression, highlighting the difficulty in estimating eQTL effects when cell type-specific gene expression levels are low.

In scenario 2, where CT1 has the highest proportion and CT3 has the lowest proportion, power to detect ct-eQTLs is reduced across models and cell types when compared with scenario 1. CSeQTL's power to detect CT1 eQTLs is much higher than OLS. CT3 eQTLs are detectable by either method if its baseline expression is high and in that case (2nd row and 2nd column of Fig. 1c) CSeQTL has much higher power than OLS, e.g., >80% power by CSeQTL vs. <20% power by OLS. OLS still has type I error inflation for CT2, to a smaller degree than in scenario 1. Finally in scenario 3, the introduction of outliers in cell type proportions substantially increases the type I error inflation of OLS for CT2 when baseline expression of CT2 is low. Additional simulation results, including results using a noisy version of cell type proportions, are presented in Supplementary Figs. 2–5.

Our implementation allows for the trimming of outliers whose Cook's distance is larger than a threshold, following the approach used by DESeq2[21]. The Cooks' distance is calculated based on the null model (no eQTL), and the value of outliers are imputed with the null model's predicted outcome (see "Methods" section "Trimming influential counts" for more details). This trimming procedure may slightly reduce power, but helps to guard against type I error. The impact of trimming is more apparent in one dataset (GTEx brain samples) in our real data analysis, which we discuss in the next section.

https://doi.org/10.1038/s41467-023-38795-w

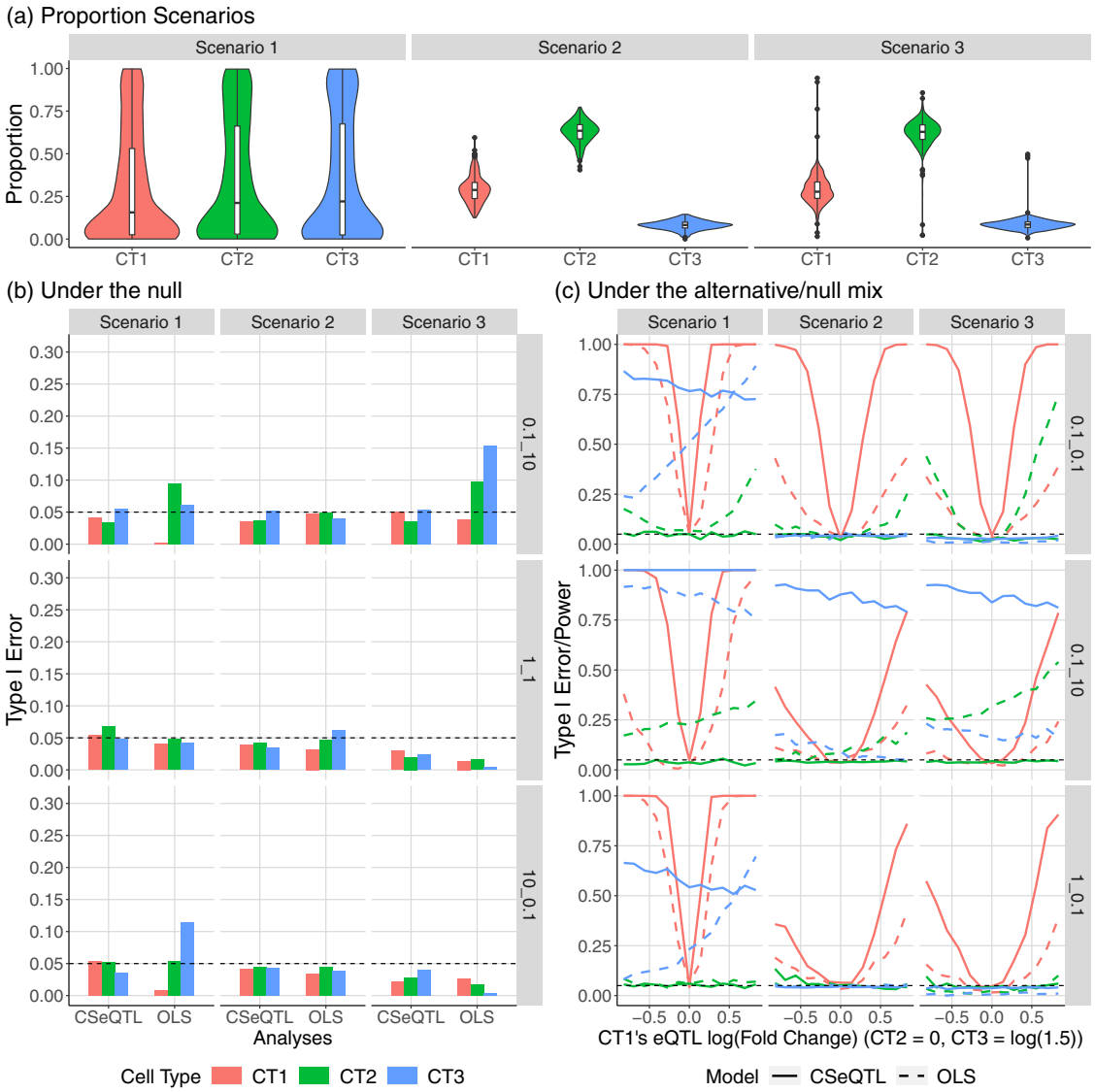

Fig. 1 | Summary of the results from simulation studies. a Simulated cell type (CT) proportions for three scenarios. Scenario 1: equally abundant and highly variable cell type proportions. Scenario 2: variable abundance and smaller variance. Scenario 3: modification of scenario 2 by adding outliers of cell type proportions. Box plots and violin plots were derived from $n = 300$ simulated cell type proportions. For each boxplot, the box ranges from Q1 (the first quartile) to Q3 (the third quartile). The median is indicated by a line across the box. The whiskers extend from Q1 and Q3 to the most extreme data points within the 1.5 IQR of the box and IQR = Q3 − Q1. **b** Simulation results under the global null (i.e., no eQTL for any cell type) for different scenarios and methods (columns of the plots) and reference allele expression configurations (rows of the plots). Each reference allele configuration is denoted by fold change of reference allele expression. For example, "10_0.1" indicates fold changes of 10 in CT2 over CT1 and 0.1 in CT3 over CT1. **c** Simulation under the mixture of null and alternative hypothesis by models, scenarios, and reference allele expression configurations. Results of (**b**) and (**c**) are obtained after trimming outliers.

In summary, power to detect ct-eQTLs is driven by the model and positively correlated with eQTL effect size, absolute and relative reference allele expression, and variability in cell type proportions.

CSeQTL identifies many more ct-eQTLs than OLS in human brain and blood

We analyzed bulk RNA-seq data from three sources: 670 whole blood samples from the Genotype Tissue Expression (GTEx) project[3], 254 schizophrenia patients and 283 controls from the CommonMind Consortium (CMC)[22,23], and 175 brain samples from GTEx. Additionally, we studied cell type-purified bulk RNA-seq data from the BLUEPRINT cohort, including purified CD4+ T cells ($n = 212$), monocytes ($n = 197$), and neutrophils ($n = 205$). For the purified bulk RNA-seq data, CSeQTL was equivalent to TReCASE, and the results were used to validate ct-eQTL results from GTEx whole blood samples.

We obtained phased genotypes, TReC, ASReC, observed covariates, and latent batch covariates for each of the four cohorts (GTEx whole blood, CMC brain, GTEx brain, and BLUEPRINT). See Supplementary Note 4 for more information. Using ICeDT[24], we estimated cell type proportions based on TReC and cell type-specific reference data for 5 brain cell types[25] and 22 blood cell types[26]. See Supplementary Note 2 for more details.

We found that the distributions of cell type proportions were similar between schizophrenia patients and healthy controls in CMC and GTEx brain samples (Fig. 2a). Excitatory neurons (Exc) had the highest proportions, followed by astrocytes (Astro), inhibitory neurons (Inh), oligodendrocytes (Oligo), and oligodendrocyte precursor cells (OPC). Microglia had the lowest proportions and the smallest variation, making it difficult to detect ct-eQTLs in this cell type. For the 22 blood cell types[26], we collapsed them into seven cell types due to limited prevalence and variability in some cell types (Supplementary

https://doi.org/10.1038/s41467-023-38795-w

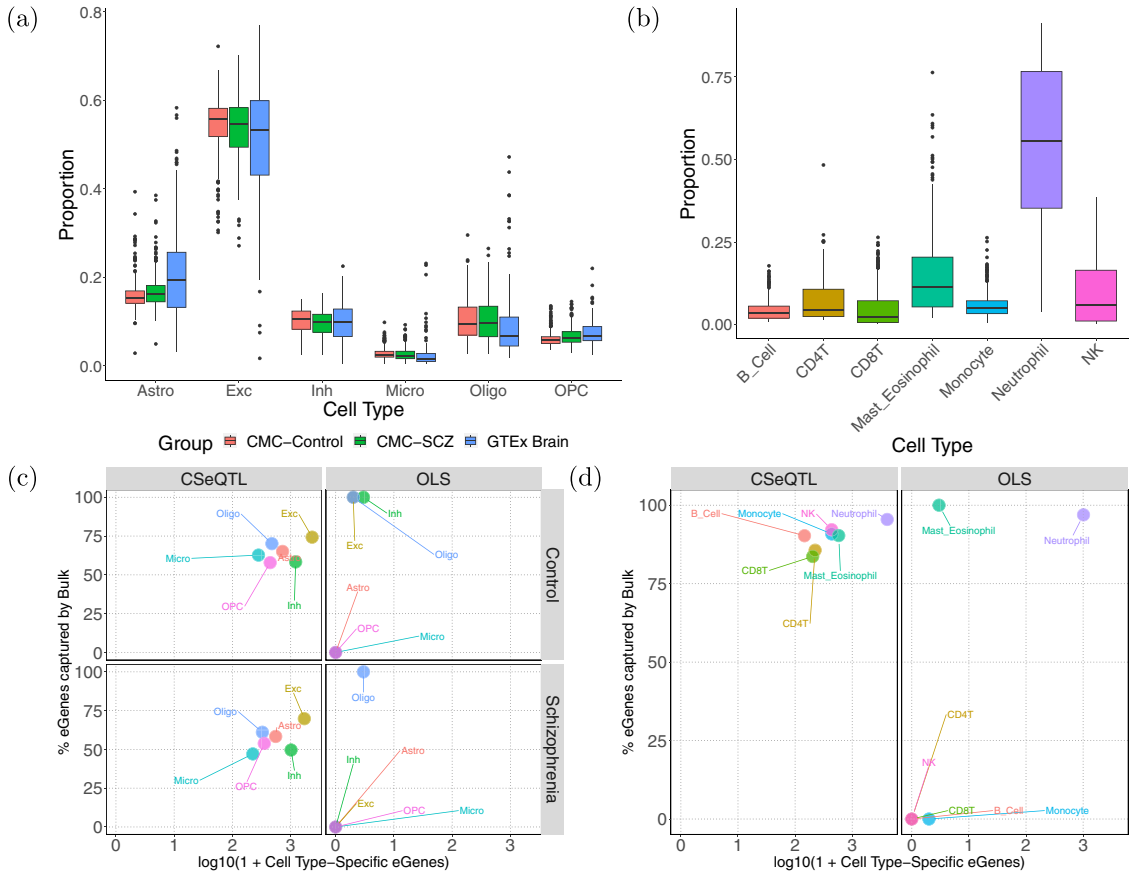

Fig. 2 | Summary of cell type compositions and the number of eGenes. a Cell type proportion estimates of six cell types astrocytes (Astro), excitatory neuron (Exc), inhibitory neuron (Inh), microglia (Micro), oligodendrocytes (Oligo), and oligodendrocyte precursor cells (OPC) from the brain samples of schizophrenia patients and controls from the CommonMind Consortium (CMC) as well as samples from GTEx Brain. Box plots are derived from $n = 283$ CMC-Control, $n = 250$ CMC-SCZ, and $n = 175$ GTEx Brain samples. **b** Cell type proportion estimates of seven cell types from whole blood samples of GTEx. Cell type proportions were first estimated for 22 cell types (Supplementary Fig. 8) and then collapsed to seven cell types to avoid individual cell types with very low proportions and variances. Box plots are derived from $n = 670$ GTEx whole blood samples. In both (**a**) and (**b**), for each boxplot, the box ranges from Q1 (the first quartile) to Q3 (the third quartile). The median is indicated by a line across the box. The whiskers extend from Q1 and Q3 to the most extreme data points within the 1.5 IQR of the box and IQR = Q3 − Q1. **c** A summary of the number of detected eGenes per stratum by method and case/control status for CMC. **d** A summary of detected eGenes by method for GTEx whole blood. For (**c**) and (**d**), the *X*-axis is the (number of cell type-specific eGenes + 1) in log (base 10) scale. The *Y*-axis denotes the percentage of cell type-specific eGenes that overlap with eGenes from bulk eQTL mapping.

Fig. 8 and "Methods" section "Grouping 22 blood cell types to seven cell types"). In GTEx whole blood samples, neutrophils were the dominant cell type with the highest proportions and the largest variance (Fig. 2b).

We conducted both ct-eQTL mapping and traditional bulk eQTL mapping, which assesses aggregated eQTL effects in bulk tissue samples. For both ct-eQTL and bulk eQTL mapping, we included a set of covariates: library size, observed covariates (such as age, sex, and known batch effects), and genotype principal components (PCs). We also added latent factors estimated from gene expression data, which were calculated by PCs of residualized gene expression data after accounting for all the aforementioned covariates. We obtained two sets of residualized TReC PCs: the first set was generated by residuals that did not account for cell type proportions and the second set did. The first set was used for bulk eQTL mapping, to mimic the common practice of eQTL mapping. The second set was used in ct-eQTL mapping. When using OLS for ct-eQTL mapping, we included cell type proportions and interaction terms between genotype and cell type proportions. We excluded genes with low expression in most samples (75th percentile of TReC <50) and SNPs with minor allele frequencies below 5% from our analysis. We considered SNPs located between 50 kilobases before the transcription start site and 50 kilobases after the transcription end site, including those within the gene body.

We trimmed expression outliers of each gene using Cook's distance. To determine the appropriate threshold of Cook's distance for each dataset, we ran TReC-only eQTL mapping using permuted data for all the genes on chromosome 1 with thresholds of 10, 15, 20, or no trimming. We selected the threshold for each dataset to ensure that type I error was controlled per cell type. The selected thresholds were 10 for GTEx brain data and 20 for the other three cohorts. A more aggressive trimming threshold was needed for GTEx brain data, likely due to its smaller sample size.

For eQTL mapping, we need to account for two layers of multiple testing: (1) testing across multiple local SNPs per gene and (2) testing across genes. For each gene, we assessed the significance of its minimum p value across all local SNPs by calculating the corresponding permutation p value. A brute-force implementation, which involves permuting the data many times and running CSeQTL on each permuted dataset, is computationally prohibitive. Instead, we used a computationally efficient method called geoP[19,27] to calculate a permutation p value by estimating the effective number of independent tests. After this step, each gene has one permutation p value. To account for multiple testing across genes, we selected a permutation p value cutoff to control false discovery rate (FDR) quantified by q-value (Supplementary Note 3). We calculated a q-value[28] for each permutation p value cutoff and chose a q-value cutoff 0.005 by default. This

cutoff was smaller than typical FDR cutoff (e.g., 0.05) because the calculation of q-value accounted for the proportion of nulls, which could lead to a liberal permutation p value cutoff when the proportion of nulls was small. For bulk eQTL results, a q-value of 0.005 corresponds to permutation p value around 0.02 while a q-value of 0.05 may correspond to a permutation p value larger than 0.1 (Supplementary Tables 7–10). We applied this two-step multiple testing correction procedure to both bulk eQTL mapping and ct-eQTL mapping for each cell type. A similar procedure has been used in GTEx studies[3].

When performing bulk eQTL mapping or ct-eQTL mapping using cell type-purified samples from BLUEPRINT, CSeQTL is equivalent to the TReCASE method[29]. Consistent with our previous results[19,29], CSeQTL has much higher power than OLS. For example, considering the results for CMC schizophrenia samples ($n = 250$), for bulk eQTL mapping after trimming outliers, CSeQTL and OLS identified around 6900 and 2900 eGenes (genes with at least one significant eQTL) respectively (Supplementary Table 2). Similar results were observed for BLUEPRINT data from purified blood cell types (Supplementary Table 4 and Supplementary Figs. 17 and 18).

CSeQTL identified many more ct-eQTLs than OLS for different brain cell types. After trimming, CSeQTL identified hundreds to thousands of ct-eQTLs per cell type in CMC schizophrenia data and OLS only identified two eQTLs in oligodendrocytes (Fig. 2c and Supplementary Table 2). The results were similar for CMC control samples ($n = 275$) and trimming outliers did not have a large impact (Supplementary Table 2). In contrast, trimming outliers had a large effect for GTEx brain data, which had a relatively small sample size of 174. In particular, for microglia, the cell type with the lowest abundance, CSeQTL and OLS identified 885 and 184 eGenes before trimming, but only 96 and 0 eGenes after trimming (Supplementary Fig. 14 and Supplementary Table 3). The results from GTEx brain data suggest that CSeQTL still has much higher power than OLS when sample size is small, but should be used with caution.

For blood ct-eQTLs estimated by GTEx whole blood data, OLS identified 1014 eGenes in neutrophil, and two or zero eGenes in other cell types. In contrast, CSeQTL identified >4000 eGenes in neutrophil, including most findings by OLS (Supplementary Fig. 21) and hundreds of eGenes in other cell types (Fig. 2d and Supplementary Table 5).

CSeQTL results demonstrated limited eGene overlaps across cell types (Supplementary Figs. 12, 15 and 20), though the majority of ct-eQTLs overlap with the eQTLs detected by bulk eQTL mapping. These results suggest ct-eQTL signals may be detectable from bulk tissue samples, though without knowing the relevant cell types. Very low consistency between ct-eQTLs and bulk eQTLs may indicate false discoveries in ct-eQTLs. For example, for the GTEx brain study, before trimming, OLS identified 1332 eQTLs in microglia for 184 eGenes, while only <0.01% overlap with bulk eQTLs, and none of these 1332 eQTLs remained significant after trimming outliers (Supplementary Table 3). In all comparisons hereafter, we focused on the eQTL results after trimming outliers since earlier results demonstrated it could reduce the number of false positives.

CSeQTL findings have significant overlaps with ct-eQTLs identified by purified bulk RNA-seq data or scRNA-seq data

We validated the CSeQTL findings from GTEx whole blood using the eQTLs identified from purified bulk RNA-seq data of three cell types −CD4T, monocyte, and neutrophils−from the BLUEPRINT project[30]. A large number of eGenes were identified from BLUEPRINT data and the number was imbalanced across cell types (Supplementary Table 6). In order to make a meaningful comparison, we compared the CSeQTL findings to the top 500 eGenes (<5% of all genes considered by BLUEPRINT) for each of the three BLUEPRINT cell types. At a q-value cutoff of 0.005 for any fold changes, around 35%, 17%, and 12% of CSeQTL eGenes from neutrophil, CD4T, and monocytes overlapped with the top 500 BLUEPRINT eGenes, respectively. These proportions increased to 40%, 30%, and 20% for q-value < 0.001 and fold change ≥1.5 (Fig. 3a, b). The numbers of overlaps were 5.7–8.8 times of the numbers expected by chance (Fig. 3c). Higher overlapping proportion in neutrophil was expected because it was the most abundant cell type and CSeQTL had higher power for the more abundant cell type.

We also compared the CSeQTL results from GTEx whole blood with the ct-eQTLs identified from a large scRNA-seq dataset[31] from peripheral blood mononuclear cells (PBMCs) of 982 donors, with an average of 1291 cells per donor. Yazar et al.[31] studied ct-eQTLs in 14 types of immune cells. We removed two cell types with very low proportions and very small number of eGenes. The remaining 12 cell types were collapsed to five categories: B, CD4T, CD8T, Monocyte, and NK (Natural Killer), matching the cell types studied in GTEx whole blood data. This was a challenging comparison because the five cell types had small proportions in whole blood samples, where the most abundant cell type was neutrophil (Fig. 2a, b). Nevertheless, we found highly significant overlaps between CSeQTL eGenes and Yazar et al. eGenes, with fold change enrichments ranging from 4.1 to 6.7 (Fig. 3d–f). The fact that more stringent criteria to select ct-eQTLs lead to larger overlap proportions (Fig. 3a, d) suggests that our quantification of ct-eQTL effect sizes and significance levels is useful to select stronger ct-eQTLs. CD4T is the most abundant cell type studied by Yazar et al.[31], though the replication percentage is lower than most other cell types, likely due to two reasons. First, its proportion is low in whole blood samples (Fig. 2b). Second, similarity between CD4T cells and other cell types, such as CD8T cells, may lead to reduced accuracy of cell type deconvolution in bulk RNA-seq data as well as cell type classification in scRNA-seq data. Another important criterion to evaluate eQTL findings is the consistency of eQTL effect directions. We examined the ct-eQTLs that were identified by both CSeQTL (using a q-value cutoff 0.1 or 0.005) and scRNA-seq data (p value < 0.01), and found the eQTL directions were consistent for more than 90% of ct-eQTL findings across most cell types (Fig. 3g). In contrast, without applying any q-value/p value filtering the consistency proportion is 51% (Supplementary Table 11).

For the ct-eQTLs identified from brain samples (GTEx brain, CMC schizophrenia patients or controls), we compared with ct-eQTLs reported by a single nucleus RNA-seq (snRNA-seq) study[32]. Bryois et al.[32] collected snRNA-seq data for 6940 to 14,595 genes in 144 to 192 individuals for eight major brain cell types: excitatory neurons, inhibitory neurons, astrocytes, microglia, oligodendrocytes, oligodendrocyte precursor cells (OPCs), Endothelial, and Pericytes. Both Pericytes and Endothelial had very small number of cells and ct-eQTLs, and thus we skipped them in our comparison. The remaining six cell types were exactly the same as the cell types considered in our CSeQTL analysis. Overall the results were consistent with the findings for immune cell types. The CSeQTL eGenes had significant overlap with the top eGenes reported by Bryois et al. (Supplementary Fig. 22). Though the overlap was low for two cell types: inhibitory neurons and microglia, likely due to low proportions of these two cell types. In addition, cell type-specific expression were similar between excitatory neurons and inhibitory neurons (Supplementary Fig. 9), which could further increase the difficulty to map ct-eQTLs for inhibitory neurons. The eQTL effect direction estimates by CSeQTL and snRNA-seq were highly consistent for most cell types except for inhibitory neurons, again suggesting that ct-eQTL mapping was challenging for this cell type (Fig. 3h and Supplementary Tables 12–14).

We further compared the ct-eQTLs identified only by scRNA-seq/snRNA-seq data or only by CSeQTL on bulk RNA-seq data. The scRNA-seq-only ct-eQTLs tended to have smaller effect sizes and larger p values. Therefore CSeQTL may have missed those scRNA-seq-only ct-eQTLs because of their weaker effects (Supplementary Fig. 23). CSeQTL combines deconvolution of gene expression and eQTL mapping into one step which accounts for the uncertainty of

https://doi.org/10.1038/s41467-023-38795-w

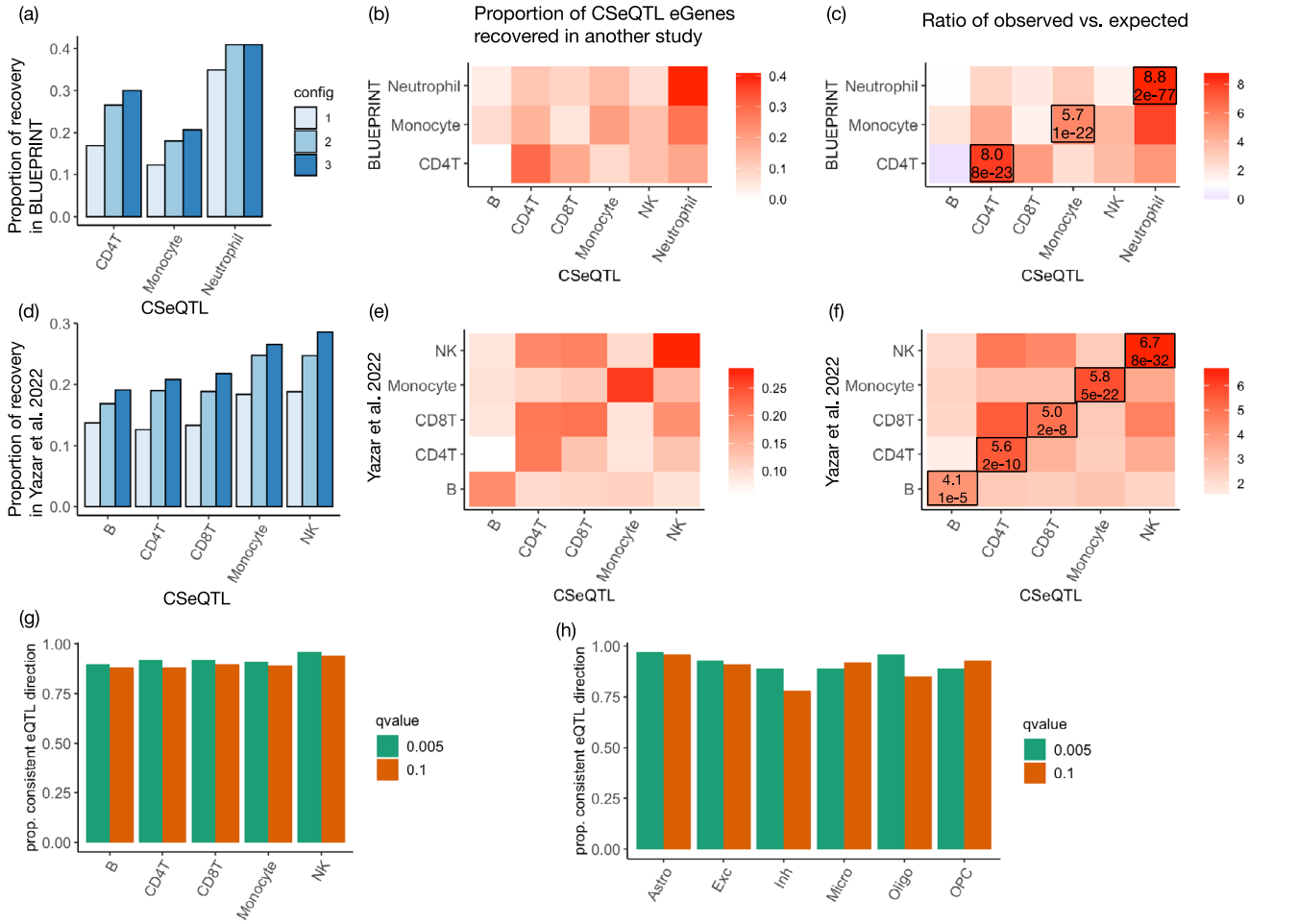

Fig. 3 | Validation of CSeQTL results. We compare the results by CSeQTL vs. the results from cell type purified bulk RNA-seq data (BLUEPRINT)[30] or scRNA-seq data from blood[31] or brain[32]. **a**, **d** The proportion of CSeQTL eGenes recovered from the top 500 (<5%) eGenes of other studies, using three configurations: (1) q-value < 0.005; (2) q-value < 0.005 and fold change ≥1.5; and (3) q-value < 0.001 and fold change ≥1.5. **b**, **e** Illustration of the recovered eGene proportions at configuration 3 for all cell type pairs. **c**, **f** For each pair of cell types studied by CSeQTL vs. BLUEPRINT or Yazar et al., we evaluated the ratio of the observed number of overlapping eGenes vs. its expected value. The ratios and corresponding p values by two-sided Fisher's exact test are labeled for each pair of matched cell types. **g**, **h** The proportion of eQTLs that have consistent directions by comparing CSeQTL results (top eQTL per gene with q-value cutoff of 0.1 or 0.005) vs. two scRNA eQTL studies with p value cutoff of 0.01.

deconvolution. Therefore, the power of CSeQTL is impacted by both the uncertainty of gene expression deconvolution and the magnitude of the ct-eQTLs. The ct-eQTLs identified solely by CSeQTL tended to have smaller effect sizes and higher expression levels in bulk samples (Supplementary Fig. 24). This makes sense because genes with higher expression have smaller uncertainty in gene expression deconvolution. Higher gene expression level should also increase the power of eQTL mapping using scRNA-seq data, though its effect could be more pronounced for CSeQTL as it also improves the accuracy of cell type deconvolution.

Characterization of ct-eQTLs

We have analyzed two blood RNA-seq datasets (BLUEPRINT and GTEx whole blood) and three brain RNA-seq datasets (CMC schizophrenia (SCZ), CMC control, and GTEx brain). It is interesting to study the consistency of eQTL results across datasets. Overall, CSeQTL results showed a higher level of consistency than OLS results (Supplementary Fig. 25 and Supplementary Table 6). For whole blood, a higher consistency was observed for neutrophil, likely due to its higher abundance. For brain datasets, CMC-SCZ and CMC-Control showed higher levels of consistency than between CMC dataset and GTEx brain, likely due to batch effects between the two studies.

We summarized the locations of the minimum p value SNPs (minP-SNPs) relative to the corresponding eGenes (Supplementary Figs. 11, 13, 16 and 19). In the brain datasets, The locations of minP-SNPs from bulk eQTL mapping showed enrichment around the transcription start site (TSS) or transcription end site (TES), though such patterns were not as clear in ct-eQTLs. A potential reason was that the eQTLs around TSS and TES were more likely to be shared across cell types. In GTEx brain results, more ct-eQTLs tended to be located further away from the corresponding eGenes, which might be due to the limited sample size hence higher uncertainty to locate the eQTLs. For the three purified cell types from BLUEPRINT (Supplementary Fig. 16), the enrichment of eQTLs around TSS was stronger than TES. Similar patterns of eQTL locations were observed for the same three cell types in GTEx whole blood samples (Supplementary Fig. 19).

Next we focused on CSeQTL results and evaluated the distribution of ct-eQTLs with respect to functional annotations of genomic regions (e.g., enhancers, promoters, 3′ UTR, 5′ UTR, etc.) by Torus[33] (Supplementary Figs. 26 and 27). For brain tissues, the functional enrichment of eQTLs for excitatory neuron, which was the most abundant cell type, was similar to the functional enrichment of bulk eQTLs. The lack of significant functional enrichment in other cell types could be partially due to smaller number of ct-eQTLs. Comparing brain samples of

https://doi.org/10.1038/s41467-023-38795-w

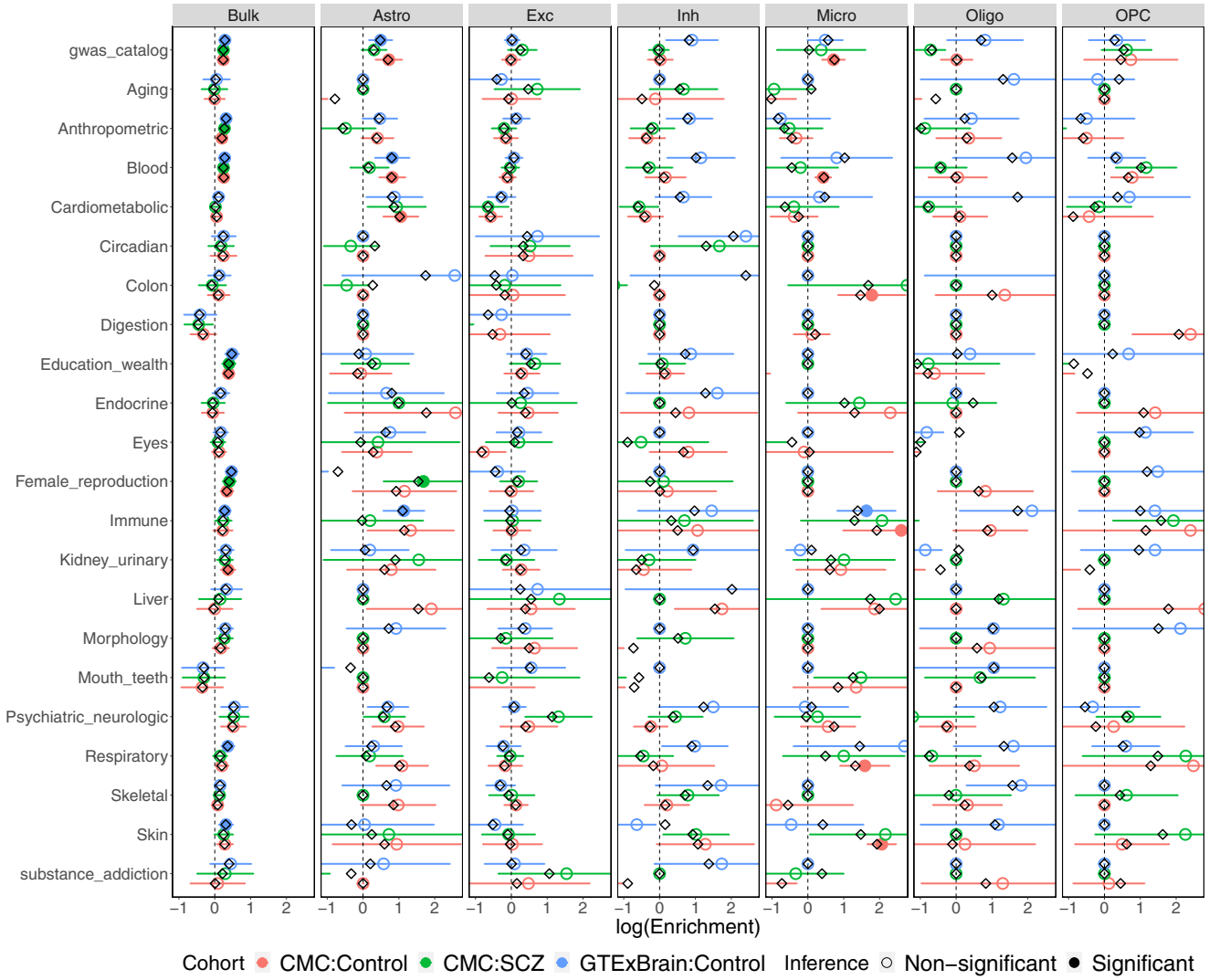

Fig. 4 | GWAS enrichment for CMC and GTEx brain. Black diamonds correspond to point estimates of log enrichment of eQTLs among GWAS hits, while open and filled circles (the centers of error bars) correspond to jackknife estimates of log enrichment. The block jackknife-based 95% confidence intervals are derived from sorting and grouping genes and loci into $n = 200$ blocks. Intervals are converted to nominal p values that are then Bonferroni corrected. Filled circles correspond to the ones with lower bound of confidence intervals larger than zero and adjusted p values < 0.05.

schizophrenia patients vs. controls (either CMC controls and GTEx controls), enrichment of eQTLs at 5′ UTR and non-coding (NC) transcript were observed in both control groups but were absent in schizophrenia patients. Since neutrophil was the dominant cell type in whole blood, as expected, functional enrichment in neutrophil and whole blood was highly consistent. Despite the small proportions of CD4+ T and monocyte in whole blood, CSeQTL results recovered similar functional enrichment as those observed in purified cell types from BLUEPRINT data.

CSeQTL helps interpret GWAS findings

EQTLs are often used to study the genetic basis of complex traits by examining their overlap with genetic loci identified from genome-wide association studies (GWAS). Here we systematically evaluated the overlap between ct-eQTLs and GWAS hits of either all the traits included in the GWAS catalog[34] on 21 categories of traits (Fig. 4 and Supplementary Fig. 5). We calculated the enrichment of eQTLs among GWAS hits by a log fold change (the proportion of GWAS hits that overlap with eQTLs vs. the proportion of genetic loci being eQTLs). See section C.3.3 of Vasyl et al.[19] for details on the computation of point estimates and their confidence intervals.

GWAS hits of several categories (e.g., education/wealth) were enriched in the bulk eQTLs of all three brain datasets, though the degree of enrichment (measured by log fold change in Fig. 4) was small. When considering ct-eQTLs by CSeQTL, due to the smaller number of eQTLs, the confidence to estimate enrichment was often low, which led to wider confidence intervals. Despite such limitation, we observed several interesting findings. For example, the GWAS hits of immune traits were enriched in the ct-eQTLs for microglia in CMC controls and GTEx brain samples, but not in CMC SCZ samples, suggesting potential SCZ-specific and ct-eQTL signals.

Blood is arguably the most accessible tissue and thus molecular biomarkers (e.g., cell type-specific gene expression) in blood can be very valuable to understand the mechanism that connects genetic variants and complex traits. Our ct-eQTL results provided a useful resource for such studies (Fig. 5). For example, enrichment of respiratory and skin disease GWAS signals among B cell specific eQTLs, and the association between liver disease GWAS hits and the eQTLs in CD8+ T cells. Earlier studies have reported that CD8+ T cells were associated with liver damage, hepatitis, immunopathology, and liver cancer[35–39].

https://doi.org/10.1038/s41467-023-38795-w

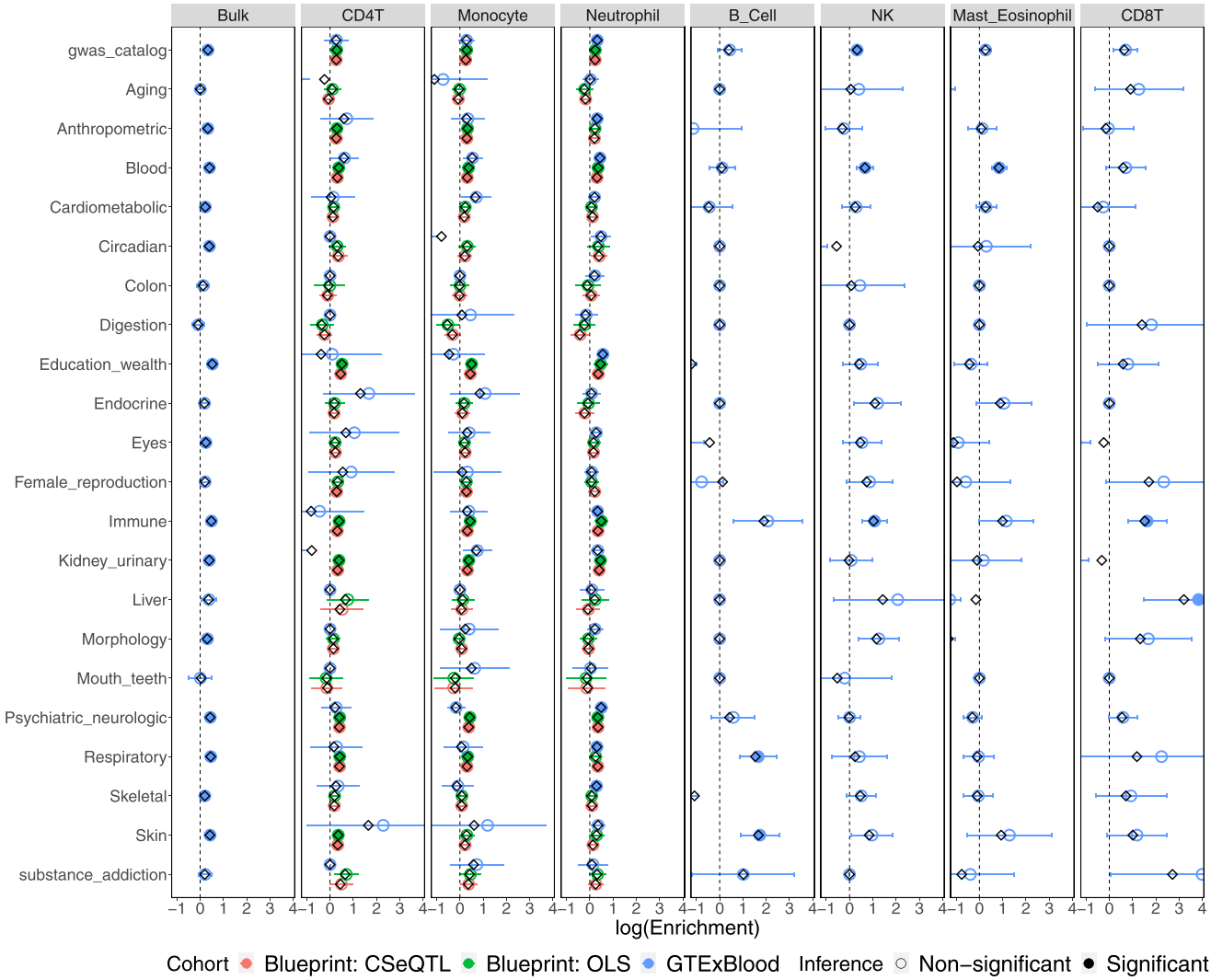

Fig. 5 | GWAS enrichment for BLUEPRINT and GTEx whole blood. Black diamonds correspond to point estimates of log enrichment of eQTLs among GWAS hits, while open and filled circles (the centers of error bars) correspond to jackknife estimates of log enrichment. The block jackknife-based 95% confidence intervals are derived from sorting and grouping genes and loci into $n = 200$ blocks. Intervals are converted to nominal p values that are then Bonferroni corrected. Filled circles correspond to the ones with lower bound of confidence intervals larger than zero and adjusted p values < 0.05.

Discussion

CSeQTL's framework allows mapping ct-eQTLs using bulk RNA-seq data, by jointly modeling the effects of cell type composition and ct-eQTLs. We have shown by simulations and real data analyses that CSeQTL can have substantially higher power than a linear regression approach, while still maintaining type I error control. This is due to the underlying statistical model of CSeQTL. Deconvolution of gene expression to individual cell types should be performed using untransformed count data[40], while eQTL mapping is often done using transformed gene expression data (e.g., normal quantile transformation) to avoid the impact of outliers in count data. Such outliers are often due to a strong positive associations between mean value and variance of count data. Our CSeQTL method satisfies these two restrictions by directly modeling count data using a negative binomial distribution that accounts for the strong mean-variance dependence. In contrast, the linear regression approach uses transformed gene expression data and models ct-eQTLs by adding interactions between cell type compositions and genotypes. The transformation of gene expression data distorts their associations with cell type compositions and thus can reduce power and inflate type I error. In addition, we also include allele-specific expression in our model to boost the power to detect ct-eQTLs.

Model optimization of CSeQTL is very challenging because the model may not be identifiable, for example, due to a lack of variation of one cell type's abundance across individuals or very low expression of one gene in one or more cell types. A naive implementation may result in sub-optimal solutions due to non-invertible observed information matrices, negative variances, or extreme parameter estimates, which can have a profound impact on hypothesis testing. We have developed a comprehensive set of assessments to ensure a robust and optimal solution is obtained. In addition, both linear regression and CSeQTL can be sensitive to outliers, and we addressed this issue by trimming those outliers based on the null model without eQTL effects. As shown in our real data analyses, such trimming can be particularly helpful for studies with smaller sample sizes.

Our applications toward human brain and blood bulk RNA-seq data demonstrate that the linear regression method often identifies none or a few ct-eGenes (with the exception of neutrophil in whole blood) while CSeQTL can identify hundreds to thousands of cell type-specific eGenes. When examining the overlap between these ct-eQTLs and GWAS findings, we have identified several interesting results but with high uncertainty in many cases. Future independent studies and comparisons with larger sample sizes may be needed to reach more definite conclusions.

A limitation when applying our method or any ct-eQTL mapping method on bulk RNA-seq data is accurate estimation of cell type composition, which in turn requires accurate cell type-specific gene expression reference. Here we have applied our method on the bulk RNA-seq data from brain and blood because these two tissues have readily available cell type-specific gene expression reference. We expect that in the near future, with the advance of the human cell atlas[1] or other similar projects, such resources will become available in more tissue types.

Our work also enables a new type of study design to jointly model scRNA-seq and bulk RNA-seq data to study ct-eQTLs. For example, scRNA-seq data can be collected in a small number of individuals, to be used as reference for cell type-specific expression. In addition, scRNA-seq data can also be used for eQTL mapping. After clustering and identification of cell types, scRNA-seq data can be converted to pseudo-bulk data of individual cell types and be used for eQTL mapping, e.g., by applying our TReCASE method[29]. The likelihood function of the TReCASE model can be combined with CSeQTL model in order to combine bulk RNA-seq and scRNA-seq data for ct-eQTL mapping. Adding scRNA-seq data to bulk RNA-seq data can alleviate some challenges when using bulk RNA data, e.g., limited variability in cell type abundance for one cell type. Adding bulk RNA-seq data to scRNA-seq data can reduce the cost, increase sample size, and avoid distortion of gene expression in the process of isolating single cells.

Methods

Statistical models

Notations and the joint model of TReC and ASReC. Since our model is the same for any gene-SNP pair, we omit gene and SNP indices to simplify notations. We use i and q as indices for sample and cell type, respectively, where $i = 1, \ldots, n$, $q = 1, \ldots, Q$, and n and Q denotes sample size and the number of cell types, respectively. Let T_i and N_i be the total read count (TReC) and the allele-specific read count (ASReC) mapped in the ith sample. Each SNP of interest has two alleles, A and B. Each gene has two haplotypes that are arbitrarily defined as haplotype 1 and 2. Let $N_i - N_{i2}$ and N_{i2} denote the ASReC mapped to the first and second haplotypes of sample i, respectively.

Let Z_i denote the phased genotype for the SNP in sample i, which takes values AA, AB, BA, or BB. Let $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ be a p-vector of baseline covariates (excluding the intercept), where T denotes vector or matrix transpose. Among the baseline covariates in our model, we adjust for log-transformed read depth, defined as the log of the 75th percentile of a sample's gene-level TReCs, a more robust measurement of read-depth than summing over all TReC values. Let ρ_{iq} denote the cell type proportion in the ith sample and qth cell type such that $\sum_{q=1}^{Q}\rho_{iq} = 1$ and $\boldsymbol{\rho}_i = (\rho_{i1}, \ldots, \rho_{iQ})^{\mathrm{T}}$. The cell type corresponding to $q = 1$ is referred to as the reference cell type. Our model is based on the following factorization:

$$P(T_i, N_i, N_{i2} | Z_i, X_i, \boldsymbol{\rho}_i) = P(T_i | Z_i, X_i, \boldsymbol{\rho}_i) P(N_i | T_i, Z_i, X_i, \boldsymbol{\rho}_i) P(N_{i2} | T_i, N_i, Z_i, X_i, \boldsymbol{\rho}_i)$$

Each factor is defined as follows:

- $P(T_i | Z_i, X_i, \boldsymbol{\rho}_i)$: given $(Z_i, X_i, \boldsymbol{\rho}_i)$, T_i is assumed to follow a negative binomial distribution with mean $\mu_i = E[T_i | Z_i, X_i, \boldsymbol{\rho}_i]$ and dispersion parameter ϕ such that $V[T_i | Z_i, X_i, \boldsymbol{\rho}_i] = \mu_i + \phi\mu_i^2$. This likelihood term corresponds to the TReC model.
- $P(N_i | T_i, Z_i, X_i, \boldsymbol{\rho}_i)$: this term describes the total number of allele-specific reads as a function of TReC. It is determined by the number of heterozygous SNPs within the gene and is a constant with respect to the parameters of eQTLs. Thus it is factored out from the likelihood.
- $P(N_{i2} | T_i, N_i, Z_i, X_i, \boldsymbol{\rho}_i)$: given $(N_i, Z_i, \boldsymbol{\rho}_i)$, the read count N_{i2} is assumed to be independent of (T_i, X_i) and follows a beta-binomial distribution with parameter π_i, which is the expected proportion of ASReC from the haplotype harboring the B allele for heterozygous samples among N_i allele-specific reads, and a dispersion parameter ψ. This likelihood term corresponds to the ASReC model.

The above likelihood framework is the same as our TReCASE method that combine TReC and ASE to map *cis*-eQTLs[20,29,41]. Similar to TReCASE, we reduce the negative binomial and beta-binomial distribution to Poisson and binomial distribution, respectively, when the data does not support a non-zero overdispersion parameter. Next we describe how to extend each component of the likelihood function for cell type-specific eQTL mapping.

Let $\mu_{i,z,q}$ be the expected TReC for the ith sample, zth phased genotype, and qth cell type. We assume a multiplicative model $\mu_{i,z,q} = \mu_{z,q}\exp\{X_i^{\mathrm{T}}\boldsymbol{\beta}\}$ and that the effect of baseline covariates $\boldsymbol{\beta}$ are the same for all cell types. We assume that the gene expression for each genotype is the summation of allelic expressions such that $\mu_{AA,q} = \mu_{A,q} + \mu_{A,q} = 2\mu_{A,q}$, $\mu_{AB,q} = \mu_{A,q} + \mu_{B,q} = \mu_{BA,q}$, and $\mu_{BB,q} = \mu_{B,q} + \mu_{B,q} = 2\mu_{B,q}$ where $\mu_{A,q}$ and $\mu_{B,q}$ denote the expected TReC for A and B alleles of the qth cell type, respectively. Define $\kappa_q = \mu_{A,q}/\mu_{A,1}$ and $\eta_q = \mu_{B,q}/\mu_{A,q}$ where $\kappa = (\kappa_1, \ldots, \kappa_Q)^{\mathrm{T}}$ and $\eta = (\eta_1, \ldots, \eta_Q)^{\mathrm{T}}$. κ_q, which is a nuisance parameter, is the fold change of the A allele's expression in the qth cell type vs. the first cell type. η_q is the eQTL effect size: the expression fold change of the B allele vs. A allele for the qth cell type.

Linear model. To establish a baseline for comparison and mimicking published analyses, we propose fitting a linear model by ordinary least squares (OLS). Let G_i denote the number of B alleles for a given phased SNP for the ith sample (i.e.,. $G_i = 0$, 1, 1, and 2 for $Z_i = AA$, AB, BA, and $Z_i = BB$, respectively). For a gene-SNP pair, the cell type-specific linear model is:

$$E[\bar{T}_i] = \zeta_0 + \sum_{j=1}^{p} X_{ij}\zeta_j + G_i\zeta_g + \sum_{q=2}^{Q}\rho_{iq}\gamma_q + G_i\sum_{q=2}^{Q}\rho_{iq}\delta_q$$

where $\bar{T}_i$ is the inverse normal quantile transformation of read-depth adjusted T_i. The benefit of the transformation is guarding against outliers on the count scale, and it is a popular choice in eQTL studies[3,42]. From the above model, we can test $H_0 : \zeta_g + \delta_q = 0$ to assess the strength of cell type-specific eQTL for the qth cell type, where $q = 2, \ldots, Q$, and test $H_0 : \zeta_g = 0$ for the reference cell type's eQTL.

TReC model. Let $\eta_q^{(\mathrm{T})}$ be the eQTL effect size for the TReC model, where the superscript $^{(\mathrm{T})}$ indicates TReC model. Given the above notations and parameters, let $\mu_{i,AA}$ be the expected TReC for the ith sample with AA genotype and it is defined such that:

$$\begin{aligned}\log(\mu_{i,AA}) &= \log\left(\sum_{q=1}^{Q}\rho_{iq}\mu_{i,AA,q}\right) = X_i^{\mathrm{T}}\boldsymbol{\beta} + \log\left(\sum_{q=1}^{Q}\rho_{iq}\mu_{AA,q}\right)\\ &= \log(2\mu_{A,1}) + X_i^{\mathrm{T}}\boldsymbol{\beta} + \log\left(\sum_{q=1}^{Q}\rho_{iq}\kappa_q\right)\end{aligned} \quad (1)$$

With similar derivations for genotypes AB, BA, and BB, we have:

$$\log(\mu_i) = \begin{cases}\log(\mu_{i,AA}) & Z_i = AA\\ \log(\mu_{i,AA}) + \log\left(1 + \xi_i^{(\mathrm{T})}\right) - \log(2) & Z_i = AB, BA\\ \log(\mu_{i,AA}) + \log\left(\xi_i^{(\mathrm{T})}\right) & Z_i = BB\end{cases}$$

where $\xi_i^{(\mathrm{T})} = \dfrac{\sum_{q=1}^{Q}\rho_{iq}\eta_q^{(\mathrm{T})}\kappa_q}{\sum_{q=1}^{Q}\rho_{iq}\kappa_q}$.

It is crucial to notice that $\xi_i^{(\mathrm{T})}$ represents the bulk eQTL effect size for the ith sample. If eQTL effect is the same across cell types ($\eta_1^{(\mathrm{T})} = \cdots = \eta_Q^{(\mathrm{T})} = \eta^{(\mathrm{T})}$), $\xi_i^{(\mathrm{T})} = \eta^{(\mathrm{T})}$ simplifying CSeQTL's TReC model to

the bulk TReC model presented by Sun[29]. For $Q=2$, CSeQTL's TReC model would correspond to pTReCASE's TReC model[20]. After centering continuous covariates among X_i and setting categorical covariates among X_i to their reference level, the intercept term of the above model represents the log-transformed expected TReC of the reference cell type with genotype *AA*. This straightforward interpretation of CSeQTL is a crucial feature for model optimization and parameter estimate interpretation.

ASReC model. Let $\eta_q^{(A)}$ be the eQTL effect size associated with the ASReC model. Let $P_{BB}(N_1|N;\pi;\psi)$ be the beta-binomial density for observing N_1 successes among N trials with success probability π and overdispersion parameter ψ. For a gene-SNP pair, the ASReC likelihood is defined as:

$$P(N_{i2}|N_i,Z_i) = \begin{cases} 1 & \text{if } N_i=0 \\ P_{BB}\left(N_{i2}|N_i;\pi_i=\xi_i^{(A)}/\left(1+\xi_i^{(A)}\right);\psi\right) & \text{if } N_i>0, Z_i=AB \\ P_{BB}\left(N_i-N_{i2}|N_i;\pi_i=\xi_i^{(A)}/\left(1+\xi_i^{(A)}\right);\psi\right) & \text{if } N_i>0, Z_i=BA \\ P_{BB}\left(N_{i2}|N_i;\pi_i=0.5;\psi\right) & \text{if } N_i>0, Z_i=AA,BB. \end{cases}$$

Similar to the TReC model:

$$\xi_i^{(A)}=\frac{\sum_q \rho_{iq}\eta_q^{(A)}\kappa_q}{\sum_q \rho_{iq}\kappa_q},$$

which is the bulk eQTL effect size estimated from ASReC. If $N_i=0$, the ASReC likelihood factors out of the joint model. Furthermore, while samples with genotypes *AA* and *BB* do not add information when estimating $\eta_q^{(A)}$ and κ_q, they contribute toward estimating ψ.

***cis*/*trans* eQTL testing and eQTL testing.** Following Sun[29], the model-specific eQTL parameters $\eta_q^{(T)}$ and $\eta_q^{(A)}$ are used to formally characterize *cis* and *trans* eQTLs. By defining $\eta_q^{(A)}=\eta_q^{(T)}\alpha_q$, the qth cell type-specific eQTL being *cis* corresponds to $\alpha_q=1$ and *trans* otherwise. For *cis*-eQTLs, we use the joint model that combines TReC and ASReC/ASE models with shared cell type-specific parameter $\eta_q=\eta_q^{(A)}=\eta_q^{(T)}$. We conduct *cis*/*trans* testing per gene-SNP pair and per cell type with $H_0:\alpha_q=1$ vs. $H_A:\alpha_q\neq 1$. Let $\alpha=(\alpha_1,\ldots,\alpha_Q)^T$.

EQTL significance testing is conducted for each gene, SNP, and cell type using either the TReC model for *trans*-eQTL with $H_0:\eta_q^{(T)}=1$ vs. $H_A:\eta_q^{(T)}\neq 1$ or the joint model for *cis*-eQTL with $H_0:\eta_q=1$ vs. $H_A:\eta_q\neq 1$. Thus our model formulation is flexible enough to allow subsets of cell type-specific eQTLs to be *cis*- or *trans*-eQTLs.

Optimization scheme and parameter assessment

Given cell type proportions, our optimization scheme for TReC and ASReC model fitting and hypothesis testing is based on the following procedure for a gene-SNP pair. This scheme helps to avoid local optima since parameter estimation can be sensitive to initialization and influential counts. Let $\theta=(\mu_{A,1},\boldsymbol{\beta}^T,\phi,\kappa^T,\eta^T,\psi,\alpha^T)^T$ denote the pre-established set of unconstrained parameters to optimize over. First, we condition T_i on X_i to obtain $\hat{\theta}_1=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T)^T$ by fitting a Poisson model with Newton-Raphson. Second, we can fit a negative binomial model with initialization $\hat{\phi}=1$ to obtain $\hat{\theta}_2=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi})^T$, also with Newton-Raphson. Third, we incorporate $\boldsymbol{\rho}_i$, initialize $\hat{\kappa}_q=1$, and use Broyden–Fletcher–Goldfarb–Shanno (BFGS) to obtain $\hat{\theta}_3=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi},\hat{\kappa}^T)^T$. Fourth, we incorporate Z_i, initialize $\hat{\eta}_q=1$, and use BFGS to obtain $\hat{\theta}_4=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi},\hat{\kappa}^T,\hat{\eta}^T)^T$. Fifth, we incorporate (N_i,N_{i2}), initialize $\hat{\psi}=1$, fix $\hat{\alpha}_q=1$, and run BFGS to obtain $\hat{\theta}_5=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi},\hat{\kappa}^T,\hat{\eta}^T,\hat{\psi})^T$. Lastly, we optimize over the full parameter set to obtain $\hat{\theta}=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi},\hat{\kappa}^T,\hat{\eta}^T,\hat{\psi},\hat{\alpha}^T)^T$, also with BFGS.

This optimization scheme does have inherent challenges to achieve stable convergence. One key regularity condition is that the estimated parameters are not on the boundary of the parameter space, in our case, $\mu_{A,q}>0$ and $\mu_{B,q}>0$ corresponding to non-zero expression for each allele and cell type. A second requirement is sufficient variability in ρ_{iq} across samples to estimate κ_q, η_q, and α_q. This is comparable to an identifiability condition for a linear regression where each covariate has non-zero variance. It is likely that these two requirements are not satisfied for some genes or cell types. Therefore the full model or set of estimable parameters needs to be adjusted. Let $l_n(\theta)$, $\dot{l}_n(\theta)$, and $\ddot{l}_n(\theta)$ denote the log-likelihood, score, and (negative) observed information, respectively. Let $|\cdot|_2$ denote the L_2 norm. Convergence is defined when $|\dot{l}_n(\hat{\theta})|_2<\epsilon_1$, $\ddot{l}_n(\theta)$ is invertible, no negative variances, and $|\ddot{l}_n(\hat{\theta})^{-1}\dot{l}_n(\hat{\theta})|_2<\epsilon_2$ for predefined thresholds ϵ_1 and ϵ_2. By default, $\epsilon_1=10^{-3}$ and $\epsilon_2=10^{-6}$. To determine which cell type-specific parameters to constrain to their null values ($\kappa_q=0$, $\eta_q=1$, $\alpha_q=1$), we run the above optimization procedure and set the unidentifiable parameters to their null values and re-run the optimization procedure, and iterate this procedure until all the remaining parameters are estimable. More specifically, we initialize our parameters with $\hat{\theta}_2$ and perform the following operations:

- First, we estimate $\hat{\theta}_2=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi})^T$ by maximum likelihood estimate (MLE), while ignoring eQTL and cell type composition.
- Next we estimate $\hat{\theta}_3=(\hat{\mu}_{A,1},\hat{\boldsymbol{\beta}}^T,\hat{\phi},\hat{\kappa}^T)^T$. The κ_q parameters are estimated relative to the reference cell type ($q=1$) and therefore we must ensure the reference cell type has non-zero TReC ($\hat{\mu}_{A,1}>0$). By default we set the reference cell type to be the one with highest average proportion across samples. After estimating $\hat{\kappa}$, we can determine which cell type has highest TReC and swap that cell type to be the reference cell type. This choice of reference cell type can vary from gene to gene. It is an internal choice for the computation purpose and in the final output, all the parameters are transformed using the cell type with highest average proportion as reference. If θ_3 cannot be reliably estimated, it indicates that some κ_q's are close to 0. We calculate $\hat{\mu}_{AA,q}\equiv 2\hat{\mu}_{A,1}\hat{\kappa}_q$. If $\min_q(\hat{\mu}_{AA,q})<2$ (each haplotype expresses at least one TReC), set $\hat{\kappa}_q=0$ and $\hat{\eta}_q=\hat{\alpha}_q=1$, and then re-optimize.
- Next we estimate eQTL effects η in θ_4 and θ_5. If convergence is achieved, move on to the next step. Otherwise, the ASReCs of one or more cell types are near zero. Then we calculate $\hat{\mu}_{Aq}\equiv\hat{\mu}_{A,1}\hat{\kappa}_q$, $\hat{\mu}_{Bq}\equiv\hat{\mu}_{Aq}\hat{\eta}_q$, and $\hat{\mu}_{zq}\equiv\min(\hat{\mu}_{Aq},\hat{\mu}_{Bq})$, and variance estimate for $\log(\hat{\eta}_q)$ (optimizing over unconstrained parameters). If $0<\hat{\mu}_{zq}<1$ or η_q variance estimate is negative, set $\hat{\eta}_q=1$ and re-optimize and repeat until convergence. For each cell type where $\hat{\eta}_q=1$, set $\hat{\alpha}_q=1$ for subsequent steps.
- Next we estimate α in θ_6. If convergence is achieved, we have established the full model. Otherwise, check for variance estimates for $\log(\hat{\alpha}_q)$ that are negative and set $\hat{\alpha}_q=1$ and re-optimize. If none of the cell types α_q variances are negative and convergence is not achieved, identify the cell type with largest α_q variance and set $\hat{\alpha}_q=1$.

In general, whenever we need to re-optimize, we simply start off at the step prior to the current step, there is no need to return to $\hat{\theta}_2$ since the procedure has established the "submodel" or nested set of estimated parameters that achieved convergence. In addition, our procedure is strictly designed to assess convergence first at each step before attempting to constrain parameter estimates or looking to the

variance estimates to avoid unnecessary matrix inversions until all other criteria are met.

Trimming influential counts

Data trimming and quality control are a crucial issue associated with regression analyses[43]. Modeling observed outcomes directly risks highly influential or potential outlier data points that contribute to biased parameter estimates and inflated type I error. For eQTL analysis, if the same subset of samples were consistently identified as outlier, we could exclude them. But among post quality control samples, an analysis could involve potentially excluding different subsets of samples per gene, risking a power loss and difficulty to interpret the results. In the case of differential expression, DESeq2[21] systematically trims outlier observations whose Cook's distance is beyond a predefined cutoff based on the F-distribution. We have adopted a similar trimming approach in our eQTL analyses.

For a given gene, we characterize the influence of a sample through our definition of Cook's distance for the ith sample with:

$$C_i = \frac{1}{m}\sum_{j=1}^{n}\frac{(\hat{\mu}_{j(i)} - \hat{\mu}_j)^2}{\hat{v}_j}$$

where $m = p + Q - 1$, $\hat{\mu}_j$ is the estimated mean TReC for the jth sample, $\hat{\mu}_{j(i)}$ is the estimated mean TReC for the jth sample after excluding the ith sample, and $\hat{v}_j = \hat{\mu}_j + \hat{\phi}\hat{\mu}_j^2$, the estimated TReC variance for the jth sample. Since our TReC model is not the traditional GLM due to the sample-specific offset term $(\log(\sum_{q=1}^{Q}\rho_{iq}\kappa_q))$, we cannot directly characterize leverage. We then calculate normalized Cook's distance to put Cook's distances on the same scale across genes, denoted $\tilde{C}_i$ and characterized as:

$$\tilde{C}_i = \frac{C_i - \mathrm{med}(C_1,\ldots,C_n)}{\mathrm{mad}(C_1,\ldots,C_n)},$$

where med(...) and mad(...) denote median and median absolute deviation, respectively. We propose trimming the original TReC (T_i) if $\tilde{C}_i > c$ where c is some predefined threshold. To calculate Cook's distance, we fit CSeQTL's TReC model without adjusting for SNP since a gene can have multiple SNPs and a gene's TReC can be influential regardless of genotype. We explored the possibility of using $C_i > 4/n$ and $C_i > F(q = 0.99, m, n - m)$ as a trimming criteria however it failed to detect clear visual outliers. We decided on an appropriate threshold on $\tilde{C}_i$ by running CSeQTL on chr1 genes with permuted SNP genotypes. We tried cutoff thresholds 40, 20, 10, and 5. The largest threshold that controls the type I error was selected. Unlike the trimmed means used by DESeq2 to impute the TReC value, we impute the TReC value with the estimated TReC for a sample from CSeQTL's TReC model without SNP adjustment.

Simulation setup

We describe how the cell type proportions are simulated. In the first scenario, let $X \sim U(a, b)$ denote a random variable X sampled from a continuous uniform distribution ranging from a to b. Specifically $\rho_{iq} = \exp\{U_{iq}\}/\sum_{s=1}^{Q}\exp\{U_{is}\}$ and $U_{iq} \sim U(-4, 4)$. In the second scenario, to allow cell types to reflect observed proportions with wide and narrow ranges of proportions, we simulated ρ_{i1} from a beta distribution with shape parameters 10 and 24 (values derived based on maximum likelihood estimates from fitting a beta distribution to CMC's astrocyte cell type proportions), $\rho_{i2} = |0.85 - 0.76\rho_{i1} - 0.03\rho_{i1}^2 + \epsilon_i|$, where ϵ_i was sampled from a centered normal distribution with standard deviation 0.02, and $\rho_{i3} = 1 - \rho_{i1} - \rho_{i2}$. If $\rho_{i3} < 0$, we set it to zero and normalize the proportions across cell types. For the third scenario, proportions are first simulated under the second scenario. Next, for each cell type, the initial proportions greater than the 99% quantile were replaced by values sampled from $U(0.7, 0.9)$ while initial proportions less than the 1% quantile were replaced by values sampled from $U(0, 0.1)$. These final values are re-normalized across cell types to sum to one.

For $n = 300$, we simulate $p = 4$ baseline covariates. The first covariate is X_{i1}, which represents read-depth, is simulated by a gamma distribution with shape parameter set to 600 and rate parameter set to 100, based on empirical MLE estimates from CMC samples. X_{i2}, which represent sex, is generated by a Bernoulli distribution with success probability of 0.5. X_{i3} is generated by a continuous uniform distribution ranging from −1 to 1. X_{i4} is simulated by a standard normal distribution. These latter two variables represent arbitrarily distributed continuous covariates. Continuous covariates X_{i1}, X_{i3}, and X_{i4} are centered and scaled with zero mean and unit variance. Assuming Hardy Weinberg equilibrium, genotypes were generated using a categorical distribution with probabilities $(1 - m_A)^2, m_A(1 - m_A), m_A(1 - m_A), m_A^2$ for AA, AB, BA, BB, respectively, where m_A denotes the minor allele frequency. We set $m_A = 0.2$.

Grouping 22 blood cell types to seven cell types

The "CD4T" cell type is defined by pooling CD4 naive, CD4 memory resting, CD4 memory activated, follicular helper, regulatory T cells (Tregs) and gamma delta cells. The gamma delta T cells is indeed a different type of T cells while all other type of T cells are alpha beta T cells. However, its proportion is very low (Supplementary Fig. 8) and thus adding it to any other cell type does not lead to any noticeable change of cell type composition. Here we combine it into the CD4T category just for the convenience of implementation. The "B_Cell" cell type is the result of combining B cell naive, B cell memory and Plasma cells. CD8 T cells were not collapsed with other cell types and simply denoted "CD8T". The "Mast_Eosinophil" cell type is composed of mast cells resting, mast cells activated, dendritic cells resting, dendritic cells activated and eosinophils. The "NK" cell type comprises of natural killer cells resting and natural killer cells activated. The "Monocytes" cell type is made up of monocytes, macrophages M0, macrophages M1 and macrophages M2. The proportion of microphage cells are very low (Supplementary Fig. 8) and thus adding them to monocytes does not make substantial changes to monocyte proportions. We further discuss the algebraic interpretation and implications of combining cell types in Supplementary Note 2.5.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Our work did not generate any new data. We have used publicly available datasets. BLUEPRINT from European Genome Archive with phased SNPs derived from whole genome sequencing (EGAD00001002663) and three purified cell types of RNA-seq bam files (EGAD00001002671, EGAD00001002674, EGAD00001002675). CommonMind data were downloaded from https://www.nimhgenetics.org/resources/commonmind. GTEx data were downloaded from NHGRI AnVIL (Genomic Data Science Analysis, Visualization, and Informatics Lab-space). Brain MTG data were downloaded from Allen Brain Institute Website https://portal.brain-map.org/atlases-and-data/rnaseq/human-mtg-smart-seq. SEA-AD snRNA-seq data were downloaded from cellxgene: https://cellxgene.cziscience.com/collections/1ca90a2d-2943-483d-b678-b809bf464c30.

Code availability

The source codes for R package CSeQTL and analysis pipeline are made publicly available at the Github repositories https://github.com/pllittle/CSeQTL(https://doi.org/10.5281/zenodo.7901725), and https://github.com/pllittle/CSeQTLworkflow (https://doi.org/10.5281/zenodo.7901800), respectively.

References

1. Regev, A. et al. Science forum: the human cell atlas. *elife* **6**, e27041 (2017).
2. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
3. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
4. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
5. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
6. Skene, N. G. et al. Genetic identification of brain cell types underlying schizophrenia. *Nat. Genet.* **50**, 825 (2018).
7. Zhu, H., Shang, L. & Zhou, X. A review of statistical methods for identifying trait-relevant tissues and cell types. *Front. Genet.* **11**, 587887 (2021).
8. Wang, R., Lin, D. Y. & Jiang, Y. Epic: Inferring relevant cell types for complex traits by integrating genome-wide association studies and single-cell RNA sequencing. *PLoS Genet.* **18**, e1010251 (2022).
9. Burgess, D. J. Getting dynamic with eQTLs. *Nat. Rev. Genet.* **20**, 500–501 (2019).
10. Strober, B. J. et al. Dynamic genetic regulation of gene expression during cellular differentiation. *Science* **364**, 1287–1290 (2019).
11. Glastonbury, C. A., Alves, A. C., Moustafa, J. S. E. S. & Small, K. S. Cell-type heterogeneity in adipose tissue is associated with complex traits and reveals disease-relevant cell-specific eQTLs. *Am. J. Hum. Genet.* **104**, 1013–1024 (2019).
12. Westra, H. J. et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).
13. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
14. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
15. Donovan, M. K., D'Antonio-Chronowska, A., D'Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 1–14 (2020).
16. Aguirre-Gamboa, R. et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinform.* **21**, 1–23 (2020).
17. Patel, D. et al. Cell-type-specific expression quantitative trait loci associated with Alzheimer disease in blood and brain tissue. *Transl. Psychiatry* **11**, 250 (2021).
18. Sun, W. & Hu, Y. eqtl mapping using RNA-seq data. *Stat. Biosci.* **5**, 198–219 (2013).
19. Zhabotynsky, V. et al. eQTL mapping using allele-specific count data is computationally feasible, powerful, and provides individual-specific estimates of genetic effects. *PLoS Genet.* **18**, e1010076 (2022).
20. Wilson, D. R., Ibrahim, J. G. & Sun, W. Mapping tumor-specific expression QTLs in impure tumor samples. *J. Am. Stat. Assoc.* **115**, 1–18 (2019).
21. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
22. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
23. Hoffman, G. E. et al. CommonMind Consortium provides transcriptomic and epigenomic data for schizophrenia and bipolar disorder. *Sci. Data* **6**, 1–14 (2019).
24. Wilson, D. R., Jin, C., Ibrahim, J. G. & Sun, W. ICeD-T provides accurate estimates of immune cell abundance in tumor samples by allowing for aberrant gene expression patterns. *J. Am. Stat. Assoc.* **115**, 1055–1065 (2020).
25. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
26. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
27. Sun, W. & Wright, F. A. A geometric interpretation of the permutation p-value and its application in eQTL studies. *Ann. Appl. Stat.* **4**, 1014–1033 (2010).
28. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).
29. Sun, W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* **68**, 1–11 (2012).
30. Chen, L. et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell* **167**, 1398–1414 (2016).
31. Yazar, S. et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science* **376**, eabf3041 (2022).
32. Bryois, J. et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat. Neurosci.* **25**, 1104–1112 (2022).
33. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
34. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
35. Dudek, M. et al. Auto-aggressive CXCR6+ CD8 T cells cause liver immune pathology in nash. *Nature* **592**, 444–449 (2021).
36. Bénéchet, A. P. et al. Dynamics and genomic landscape of CD8+ T cells undergoing hepatic priming. *Nature* **574**, 200–205 (2019).
37. Wong, Y. C., Tay, S. S., McCaughan, G. W., Bowen, D. G. & Bertolino, P. Immune outcomes in the liver: is CD8 T cell fate determined by the environment? *J. Hepatol.* **63**, 1005–1014 (2015).
38. John, B. & Crispe, I. N. Passive and active mechanisms trap activated CD8+ T cells in the liver. *J. Immunol.* **172**, 5222–5229 (2004).
39. Breuer, D. A. et al. CD8+ T cells regulate liver injury in obesity-related nonalcoholic fatty liver disease. *Am. J. Physiol. Gastrointest. Liver Physiol.* **318**, G211–G224 (2020).
40. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nat. Methods* **9**, 8–9 (2012).
41. Hu, Y. J., Sun, W., Tzeng, J. Y. & Perou, C. M. Proper use of allele-specific expression improves statistical power for cis-eQTL mapping with RNA-seq data. *J. Am. Stat. Assoc.* **110**, 962–974 (2015).
42. Wright, F. A. et al. Heritability and genomics of gene expression in peripheral blood. *Nat. Genet.* **46**, 430–437 (2014).
43. Allen, M. *The SAGE Encyclopedia of Communication Research Methods* (Sage Publications, 2017).

Acknowledgements

NIH grant R01 GM105785 for P.L., S.L., V.Z., and W.S., R56 AG079291 for Y.L., and R01HG009974 for D.-Y.L.

Author contributions

W.S., D.-Y.L., and Y.L. supervised the project. P.L. and W.S. designed the method, acquired and preprocessed the four datasets used in this paper. P.L. wrote the software package to perform simulation and real data analyses. V.Z. provided the geoP software. S.L. and W.S. provided the validation analyses. P.L., S.L., and W.S. wrote the manuscript, with input from D.-Y.L., Y.L., and V.Z.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-023-38795-w.

Correspondence and requests for materials should be addressed to Paul Little or Wei Sun.

Peer review information *Nature Communications* thanks Cathal Seoighe, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.